Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols

Shanrong Zhao[1*], Zhan Ye[2], Robert Stanton[11] Integrative Biology Center of Excellence

[2] Early Clinical Development

Pfizer Worldwide Research and Development, Cambridge, MA 02139, USA

* Corresponding author: Shanrong.Zhao@pfizer.com

Abstract

In recent years RNA-sequencing (RNA-seq) has emerged as a powerful technology for transcriptome profiling. For a given gene, the number of mapped reads is not only dependent on its expression level and gene length, but also the sequencing depth. To normalize these dependencies, RPKM (Reads Per Kilobase of transcript per Million reads mapped) and TPM (Transcripts Per Million) are used to measure gene or transcript expression levels. A common misconception is that RPKM and TPM values are already normalized, and thus should be comparable across samples or RNA-seq projects. However, RPKM and TPM represent the relative abundance of a transcript among a population of sequenced transcripts, and therefore depend on the composition of the RNA population in a sample. Quite often, it is reasonable to assume that total RNA concentration and distributions is very close across compared samples. Nevertheless, the sequenced RNA repertoires may differ significantly under different experimental conditions and/or across sequencing protocols; thus, the proportion of gene expression is not directly comparable in such cases. In this review, we illustrate typical scenarios in which RPKM and TPM are misused, unintentionally, and hope to raise scientists' awareness of this issue when comparing them across samples or different sequencing protocols.

Keywords: RNA-seq, normalization, RPKM, FPKM, TPM

Introduction

In recent years, RNA-seq has emerged as a powerful technology for transcriptome profiling (Mortazavi et al. 2008; Zhao et al. 2014; Zhao et al. 2015). In 2008, Mortazavi et al. used RNA-seq to quantify transcript prevalence for the first time (Mortazavi et al. 2008). RNA-seq avoids some of the technical limitations of microarrays, including varying probe performance, cross-hybridization, nonspecific hybridization, and dynamic range issues. RNA-seq can also detect low abundance transcripts, novel transcripts, alternative splice forms of transcripts, genetic variants and gene fusions (Zhao et al. 2014; Zhang et al. 2018). Because RNA-seq does not rely on a pre-designed complementary sequence detection probe, it is not limited to the interrogation of selected probes on an array and can also be applied to species for which the whole reference genome is not yet assembled. Thus, RNA-seq delivers both less biased and previously unknown information about the transcriptome.

In a standard RNA-seq experiment, RNAs from different sources (blood, tissue, cell lines) are purified, typically enriched with oligo (dT) primers, and then fragmented. After size selection, millions or even billions of short sequence reads are generated from a randomly fragmented cDNA library (Zhao et al. 2015; Zhao et al. 2018). The major steps in RNA-seq data analysis include quality control, read alignment, quantification of gene and transcript expression levels, normalization, analysis of differential gene expression, characterization of alternative splicing, functional analysis and gene fusion detection. The algorithms and challenges associated with each step have been reviewed elsewhere (Garber et al. 2011; Conesa et al. 2016; Zhao et al. 2016). RNA-seq has a wide variety of applications in biological research, drug discovery and development (Khatoon et al. 2014). However, the most common and popular application of RNA-seq is the identification of differentially expressed genes (DEGs) or isoforms between two or more conditions. These DEGs may serve as drug targets and biomarkers for clinical diagnosis, improve our understanding of disease pathophysiology, help determining a compound's mechanism of action, and assist with patient stratification (Khatoon et al. 2014).

Measures of expression: RPKM/FPKM and TPM

In RNA-seq, the expression level of each mRNA transcript is measured by the total number of mapped fragments, which is expected to be directly proportional to its abundance level. However, after calculating the read counts, data normalization is essential to ensure accurate inference of gene expressions (Dillies et al. 2013; Li et al. 2015; Evans et al. 2018). Raw counts mapped to a given gene are not comparable between samples or conditions because the sequencing depths or library sizes (the total

number of mapped reads) typically vary from sample to sample. Raw counts of different genes within one sample are also not directly comparable, because longer transcripts have more reads mapped to them compared with shorter transcripts of a similar expression level. Therefore, instead of using integer counts directly, normalized expression units such as RPKM (Reads Per Kilobase of transcript per Million reads mapped), FPKM (Fragments Per Kilobase of transcript per Million fragments mapped), and TPM (Transcripts Per Million), are necessary to remove technical biases in sequenced data. FPKM is closely related to RPKM except with fragment (a pair of reads) replacing read (the reason for this nomenclature is historical, since initially reads were single-end, but with the advent of paired-end sequencing it now makes more sense to speak of fragments, and hence FPKM).

RPKM was initially introduced to facilitate transparent comparison of transcript levels both within and between samples, as it re-scales gene counts to correct for differences in both library sizes and gene length (Mortazavi et al. 2008). Since RPKM was introduced, it has been widely used due to its simplicity.

$$\mathrm{RPKM} = 10^9 * \frac{Reads\ mapped\ to\ the\ transcript}{Total\ reads * Transcript\ length}$$

The intended meaning of RPKM is a measure of relative RNA molar concentration (rmc) of a transcript in a sample. If a measure of RNA abundance is proportional to rmc, then their average over genes within a sample should be a constant, namely the inverse of the number of transcripts mapped. Unfortunately, RPKM does not respect this invariance property and thus cannot be an accurate measure of rmc (Wagner et al. 2012). In fact, the average RPKM varies from sample to sample. Therefore, TPM (Transcripts Per Million), a slight modification of RPKM, was proposed (Li and Dewey 2011; Wagner et al. 2012).

$$\mathrm{TPM} = 10^6 * \frac{reads\ mapped\ to\ transcript\ /\ transcript\ length}{Sum\ (reads\ mapped\ to\ transcript\ /\ transcript\ length)}$$

TPM and RPKM are closely related. It is straightforward to convert a RPKM to a TPM using the formula below.

$$\mathrm{TPM} = 10^6 * \frac{RPKM}{Sum\ (RPKM)}$$

By definition, TPM and RPKM are proportional. However, TPM is unit-less, and it additionally fulfils the invariant average criterion. For a given RNA sample, if you were to sequence one million full length transcripts, a TPM value represents the number of transcripts you would have seen for a given gene or

isoform. The average TPM is equal to 10^6 (1 million) divided by the number of annotated transcripts in a given annotation, and thus is a constant. TPM is a better unit for RNA abundance since it respects the invariance property and is proportional to the average rmc, and thus adopted by the latest computational algorithms for transcript quantification such as RSEM (Li and Dewey 2011), Kallisto (Bray et al. 2016) and Salmon (Patro et al. 2017). Therefore, TPM will be used in the subsequent discussions unless mentioned otherwise, and examples will be given to illustrate how it can be misused.

Given the utility of RPKM and TPM in comparing gene expression values within a sample, it is not surprising that researchers would also seek to use the metrics for comparisons across projects and datasets. While conceptually valid, this type of cross-sample comparison can be problematic. As TPM values are already normalized, t is easy to assume they should be comparable across samples. Unfortunately, this is not always true. In this review, we illustrate typical scenarios in which direct comparison of RPKM and TPM across samples is problematic. To demonstrate, three public datasets were downloaded from the Sequence Read Achieve (SRA) and processed with Salmon (Patro et al. 2017) using Gencode (Harrow et al. 2012) Release 29. The choices were based upon in-house evaluations of isoform quantification algorithms (Zhang et al. 2017) and different gene models (Zhao 2014; Zhao and Zhang 2015).

Sample preparation protocol can greatly affect expression values

Ribosomal RNA (rRNA) is the most highly abundant component of total RNA isolated from animal or human cells and tissues, comprising the majority (>80% to 90%) of the molecules in a total RNA sample (O'Neil et al. 2013; Fang and Akinci-Tolun 2016). To allow efficient transcript/gene detection, highly abundant rRNAs must be removed from total RNA before sequencing. Standard approaches include selection of polyadenylated RNA (polyA) transcripts using oligo (dT) primers, or depletion of rRNAs through hybridization capture followed by magnetic bead separation. However, the polyA+ selection and rRNA depletion methods each have their unique advantages and limitations. In principle, polyA+ selection mainly captures mature mRNAs with polyA tails, whereas the rRNA depletion method can sequence both mature and immature transcripts.

Both polyA+ selection and rRNA depletion were evaluated for gene quantification in clinical RNA sequencing using human blood and colon tissue samples (Zhao et al. 2018). The same samples were prepared and sequenced using both protocols. All the raw sequencing reads were deposited into the

NCBI Sequence Read Archive under the accession number SRP127360. All sequenced transcripts were broken down into five categories according to their annotated biotypes in Gencode (**Figure 1A**). For both blood and colon samples, the most abundant category with polyA+ selection was protein-coding genes, whereas in the rRNA depletion protocol it was small RNAs. As shown in **Figure 1A**, the sequenced RNA repertoires between the polyA+ selection and rRNA depletion protocols are quite different. As a result of the different sample preparation protocols, the TPM values are not directly comparable, despite that they are derived from the same sample. In the blood sample (**Figure 1B)** sequenced by the polyA+ selection, the top three genes represent only 4.2% of transcripts (HBA2:1.5%, S100A9:1.4%, and FTL:1.3%). In contrast, in the rRNA depletion, the top three genes (RN7SL2:34.3%, RN7SL1:31.4%; and RN7SK:9.3%) represent 75% of sequenced transcripts. As a result, the expression levels of many other genes are artificially deflated in the rRNA depletion sample. For the blood sample, the log2 ratio of TPM values between polyA+ selection and rRNA depletion was calculated for individual genes. The distribution of log2 ratio is depicted in **Figure 1C**, in which the mean values for protein-coding and small RNA genes are shown as dotted lines. For protein-coding genes, TPM values tend to be higher in the polyA+ selection, while for small RNAs, the tendency is exactly opposite.

The different distribution of mRNAs across tissue types can mislead comparisons

Since different tissues express diverse RNA repertoires, TPM values across tissues should not be considered directly comparable. To demonstrate this point, RNA-seq samples corresponding to six tissue types from the same subject GTEX-N7MS were downloaded from the Genotype-Tissue Expression (GTEx) project (Carithers and Moore 2015) and processed. The percentages of transcripts from mitochondria and the top three most abundant transcripts are shown in **Figure 2A**. An examination of blood and heart tissues makes the problem clear. In heart, 48.3% of sequenced transcripts are from mitochondria, while in blood this percentage drops to as low as 1.5%. Mitochondria generate most of the cell's supply of adenosine triphosphate (ATP), used as a source of chemical energy, and play an important role in the control of cell death in cardiac myocytes (Gustafsson and Gottlieb 2008). Thus, it is not surprising to see that mitochondrial genes are actively transcribed and highly expressed in heart. In heart, the top three highly expressed genes correspond to *MT-ATP6*, *MT-ATP8* and *MT-CO3*, and represent a total of 17.4% of transcripts (**Figure 2A**). In blood, the top three genes (HBA2, HBB and HBA1) constitute as high as 81.8% of sequenced transcripts. Considering the sequenced RNA repertoires differ so dramatically, direct comparison of TPM values across tissues can be misleading.

The blood transcriptome in **Figure 2A** has a high complement of globin RNA that could potentially saturate next-generation sequencing platforms, masking lower abundance transcripts. To circumvent this issue, many commercially available globin RNA reduction kits have been developed (Mastrokolias et al. 2012; Shin et al. 2014). The top three genes (HBA2, HBB and HBA1) in this blood sample constitute as high as 81.8% of sequenced transcripts. If a very effective globin reduction kit is used, all goblins are efficiently cleared. Accordingly, compared to RNA-seq without globin reduction, TPM values for the remaining genes in the same sample will increase about five-fold after globin reduction. This is another example where differences in TPM values would be due to the experimental protocol and not biologically relevant.

RNA compartmentalization affects TPM values between cytosolic and nuclear RNA-seq

The starting material for RNA-seq studies is usually total RNA or polyA+ enriched RNA. Several limitations arise from analysing these heterogeneous pools of RNA molecules from nucleus, cytoplasm and mitochondria. Although total RNA-seq has been shown to provide insight into ongoing transcription and co-transcriptional splicing in the nucleus (Tilgner et al. 2012), the simultaneous presence of mature RNAs from the cytoplasm confounds the analysis of nuclear RNA maturation steps. Thus, the RNA-seq of separated cytosolic and nuclear RNA (**Figure 2B**) can significantly improve the analysis of complex transcriptomes from mammalian tissues (Zaghlool et al. 2013). In comparison with conventional polyA+ RNA, cytoplasmic RNA contains a significantly higher fraction of exonic sequences, providing increased sensitivity in expression analysis and splice junction detection. Conversely, the nuclear fraction shows an enrichment of unprocessed RNA compared with total RNA-seq, making it suitable for analysis of nascent transcripts and RNA processing dynamics (Zaghlool et al. 2013). Considering the large differences in RNA repertoires between nucleus and cytoplasm (Tilgner et al. 2012), the direct comparison of TPM values across cellular compartments of the same sample or between samples is not recommended.

The "strandness" of RNA-seq has a substantial impact on transcriptome profiling

Non-stranded RNA-seq does not retain the strand specificity of origin for each sequencing read. Without strand information it is difficult - sometimes impossible - to accurately quantify expression levels for genes with overlapping genomic loci that are transcribed from opposite strands (Pomaznoy et al. 2019). In contrast, stranded RNA-seq retains the strand information of a read, and thus can resolve read ambiguity in overlapping genes transcribed from opposite strands to provide a more accurate quantification of gene expression levels (Zhao et al. 2015). The scatter plots of gene expression profiles

for four biological replicates of blood samples (raw data downloaded from SRA under accession SRP056985) are shown in **Figure 3**. When comparing the same samples sequenced by the non-stranded and stranded protocols, there are many genes that are poorly correlated. It is not unusual that there are genes whose expression levels are high in one protocol, but very low or even zero in the other protocol. When the stranded versus non-stranded sequencing groups were compared, as many as 1751 genes were identified to be differentially expressed (a fold change greater than 1.5 and a Benjamini-Hochberg adjusted p-value smaller than 0.05) (Zhao et al. 2015). Thus, whether an analysis uses stranded RNA-seq or not has a substantial impact on transcriptome profiling and expression measurements for many genes.

Caution on RPKM and TPM comparison across samples with varying mRNA levels

RPKM and TPM represent relative abundance of a gene or transcript in a sample. The direct comparison of RPKM and TPM across samples is meaningful only when there are equal total RNAs between compared samples and the distribution of RNA populations are close to each other. Although equal total RNAs are generally expected, it is rarely tested and not always met. For instance, cellular stress can dramatically alter the amount of RNA in cells, as shown for heat-shock treated cells (van de Peppel et al. 2003). Furthermore, a comparison of embryonic stem cells and fibroblasts revealed a 5.5-fold difference in mRNA levels (Islam et al. 2011). Additionally, it was recently found that cells with high levels of c-Myc can amplify their gene expression program, producing two to three times more total RNA and generating cells that are larger than their low-Myc counterparts (Nie et al. 2012). Thus, under both natural and experimental conditions, the critical assumption that cells produce similar levels of RNA/cell between cell types, disease states or developmental stages is not always valid. Depending on severity, these differences can influence the biological interpretation of gene expression values. RPKM and TPM represent relative abundance of transcripts in a sample but do not normalize for global shifts in total RNA contents (Aanes et al. 2014).

Discussions and Conclusions

The sequenced RNA repertoire can vary due to differences in RNA extraction & isolation protocols (total RNA-seq vs polyA+ selection), difference in library preparation protocols (stranded vs non-stranded), and RNA abundance differences in mitochondrial and nuclear RNA compartments across tissues. Such differences should be controlled prior to comparing mRNA abundances across samples, even when using

TPM normalization. Below is a suggested workflow to follow in order to compare RPKM or TPM values across samples.

1. Make sure both samples are sequenced using the same protocol in terms of strandedness. If not, samples cannot be compared.
2. Make sure both samples use the same RNA isolation approach (polyA+ selection vs ribosomal RNA depletion). If not, they should not be compared.
3. Check the fraction of the ribosomal, mitochondrial and globin RNAs, and the top highly expressed transcripts and see whether such RNAs constitute a very large part of the sequenced reads in a sample, and thus decrease the sequencing 'real estate' available for the remaining genes in that sample. If the calculated fractions in two samples differ significantly, do not compare RPKM or TPM values directly.

TPM should never be used for quantitative comparisons across samples when the total RNA contents and its distributions are very different. However, under appropriate circumstances, TPM can be still useful for qualitative comparison such as PCA and clustering analysis. In practice, it's not common to use RPKM or TPM directly in differential analysis. Instead, counts-based methods such as DESeq (Anders and Huber 2010) and edgeR (Robinson et al. 2010; Robinson and Oshlack 2010) have been developed to identify differentially expressed (DE) genes. The fundamental assumptions underlying DESeq and edgeR are summarized as follows.

1. Most genes are not DE.
2. DE and non-DE genes behave similarly.
3. Balanced expression changes, i.e. the number and magnitude of up- and down- regulated genes are comparable.

Normalization methods would perform poorly when the assumptions above are violated. RNA-seq normalization plays a crucial role to ensure the validity of gene counts for downstream differential analysis (Dillies et al. 2013; Costa-Silva et al. 2017). However, to select the right between-sample RNA-seq normalization methods for differential analysis is beyond the scope of this review, and reviewed elsewhere (Evans et al. 2018).

As more and more RNA-seq datasets are generated, meta-analyses of large-scale RNA-seq datasets are becoming increasingly common. In this review, we illustrated how easily RPKM and TPM can be unintentionally misused, resulting in misleading conclusions that can be attributed simply to technical

differences to which researchers may not be attuned. It can be reasonable to assume that the partitioning of total RNA among the different compartments (ribosomal RNA, pre-mRNA, mitochondrial RNA, genomic pre-mRNA and polyA+ RNA) of the transcriptome is comparable across samples in a given RNA-seq project. This should be a key consideration in the initial experimental design. However, cross-study analyses are frequently done without proper control for these factors. Sequenced RNA repertoires may change substantially under different experimental conditions and/or across different sequencing protocols; thus, the proportions of gene expressions are not directly comparable in such cases. Therefore, it is strongly recommended to always check whether the total RNA amount and the composition of the RNA population are close to each other when comparing RPKM/TPM values across samples and sequenced RNA repertories. Otherwise, the comparison might be misleading, or become even pointless.

Competing financial interests

All authors declare that they have no competing interests.

Author contributions

SZ conceived and designed the study. SZ, ZY and RS participated in writing the manuscript. All authors approved the final manuscript.

Acknowledgments

The authors would like to thank Ken Dower, Enoch Huang and Abby Hill for their critical reading of the draft manuscript.

References

Aanes H, Winata C, Moen LF, Ostrup O, Mathavan S, Collas P, Rognes T, Alestrom P. 2014. Normalization of RNA-sequencing data from samples with varying mRNA levels. *PLoS One* **9**: e89158.

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biol* **11**.

Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**: 525-527.

Carithers LJ, Moore HM. 2015. The Genotype-Tissue Expression (GTEx) Project. *Biopreservation and biobanking* **13**: 307-308.

Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szczesniak MW, Gaffney DJ, Elo LL, Zhang X et al. 2016. A survey of best practices for RNA-seq data analysis. *Genome Biol* **17**: 13.

Costa-Silva J, Domingues D, Lopes FM. 2017. RNA-Seq differential expression analysis: an extended review and a software tool. *PLoS One* **12**: e0190152.

Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J et al. 2013. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in bioinformatics* **14**: 671-683.

Evans C, Hardin J, Stoebel DM. 2018. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Briefings in bioinformatics* **19**: 776-792.

Fang N, Akinci-Tolun R. 2016. Depletion of Ribosomal RNA Sequences from Single-Cell RNA-Sequencing Library. *Current protocols in molecular biology* **115**: 7.27.21-27.27.20.

Garber M, Grabherr MG, Guttman M, Trapnell C. 2011. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat Methods* **8**: 469-477.

Gustafsson AB, Gottlieb RA. 2008. Heart mitochondria: gates of life and death. *Cardiovascular research* **77**: 334-343.

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S. 2012. GENCODE: the reference human genome annotation for the ENCODE project. *Genome Res* **22**.

Islam S, Kjallquist U, Moliner A, Zajac P, Fan JB, Lonnerberg P, Linnarsson S. 2011. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res* **21**: 1160-1167.

Khatoon Z, Figler B, Zhang H, Cheng F. 2014. Introduction to RNA-Seq and its applications to drug discovery and development. *Drug development research* **75**: 324-330.

Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**: 323.

Li P, Piao Y, Shon HS, Ryu KH. 2015. Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data. *BMC Bioinformatics* **16**: 347.

Mastrokolias A, den Dunnen JT, van Ommen GB, t Hoen PA, van Roon-Mom WM. 2012. Increased sensitivity of next generation sequencing-based expression profiling after globin reduction in human blood RNA. *BMC genomics* **13**: 28.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**: 621-628.

Nie Z, Hu G, Wei G, Cui K, Yamane A, Resch W, Wang R, Green DR, Tessarollo L, Casellas R et al. 2012. c-Myc is a universal amplifier of expressed genes in lymphocytes and embryonic stem cells. *Cell* **151**: 68-79.

O'Neil D, Glowatz H, Schlumpberger M. 2013. Ribosomal RNA depletion for efficient use of RNA-seq capacity. *Current protocols in molecular biology* **Chapter 4**: Unit 4.19.

Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. 2017. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* **14**.

Pomaznoy M, Sethi A, Greenbaum J, Peters B. 2019. Identifying inaccuracies in gene expression estimates from unstranded RNA-seq data. *Scientific reports* **9**: 16342.

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**.

Robinson MD, Oshlack A. 2010. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* **11**: R25.

Shin H, Shannon CP, Fishbane N, Ruan J, Zhou M, Balshaw R, Wilson-McManus JE, Ng RT, McManus BM, Tebbutt SJ. 2014. Variation in RNA-Seq transcriptome profiles of peripheral whole blood from healthy individuals with and without globin depletion. *PLoS One* **9**: e91041.

Tilgner H, Knowles DG, Johnson R, Davis CA, Chakrabortty S, Djebali S, Curado J, Snyder M, Gingeras TR, Guigo R. 2012. Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Res* **22**: 1616-1625.

van de Peppel J, Kemmeren P, van Bakel H, Radonjic M, van Leenen D, Holstege FC. 2003. Monitoring global messenger RNA changes in externally controlled microarray experiments. *EMBO reports* **4**: 387-393.

Wagner GP, Kin K, Lynch VJ. 2012. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci* **131**: 281-285.

Zaghlool A, Ameur A, Nyberg L, Halvardson J, Grabherr M, Cavelier L, Feuk L. 2013. Efficient cellular fractionation improves RNA sequencing analysis of mature and nascent transcripts from human tissues. *BMC biotechnology* **13**: 99.

Zhang C, Dower K, Zhang B, Martinez RV, Lin LL, Zhao S. 2018. Computational identification and validation of alternative splicing in ZSF1 rat RNA-seq data, a preclinical model for type 2 diabetic nephropathy. *Scientific reports* **8**: 7624.

Zhang C, Zhang B, Lin L-L, Zhao S. 2017. Evaluation and comparison of computational tools for RNA-seq isoform quantification. *BMC genomics* **18**: 583.

Zhao S. 2014. Assessment of the impact of using a reference transcriptome in mapping short RNA-Seq reads. *PLoS One* **9**: e101374.

Zhao S, Fung-Leung WP, Bittner A, Ngo K, Liu X. 2014. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PLoS One* **9**: e78644.

Zhao S, Zhang B. 2015. A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification. *BMC genomics* **16**: 97.

Zhao S, Zhang B, Gordon W, Zhang Y, Du S, Paradis T, Vincent M, Von Schack D. 2016. Bioinformatics for RNA-Seq data analysis. In *Bioinformatics - Updated Features and Applications*, pp. 125-149, InTech.

Zhao S, Zhang Y, Gamini R, Zhang B, von Schack D. 2018. Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion. *Scientific reports* **8**: 4781.

Zhao S, Zhang Y, Gordon W, Quan J, Xi H, Du S, von Schack D, Zhang B. 2015. Comparison of stranded and non-stranded RNA-seq transcriptome profiling and investigation of gene overlap. *BMC genomics* **16**: 675.

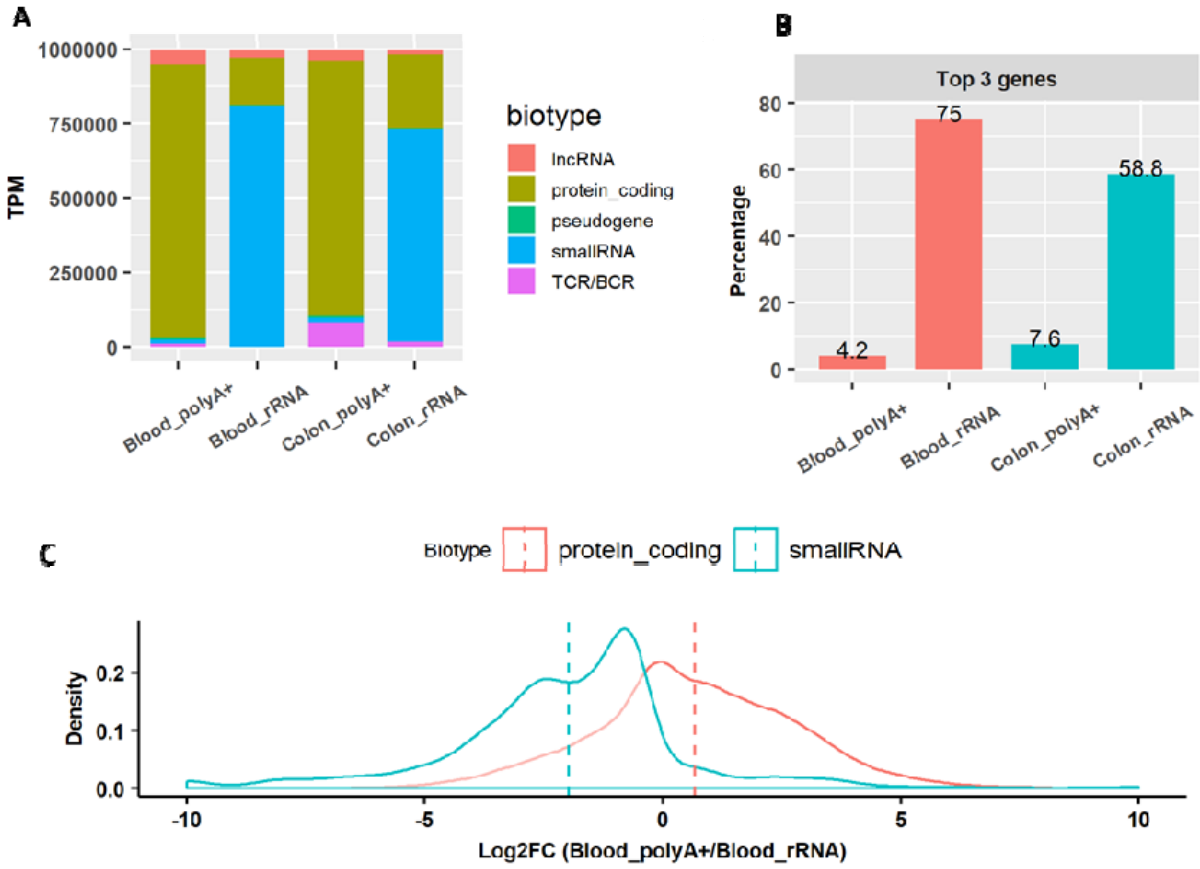

Figure 1. Comparison of TPM values of blood or colon samples with either polyA+ selection or rRNA deletion. The same blood and colon RNA samples were sequenced by both protocols (denoted as polyA+ and rRNA, respectively). A) The breakdown of sequenced transcripts by their biotype; B) The percentages of the top three highly expressed genes; and C) The distribution of log2 ratio of TPM values in polyA+ selection over rRNA deletion.

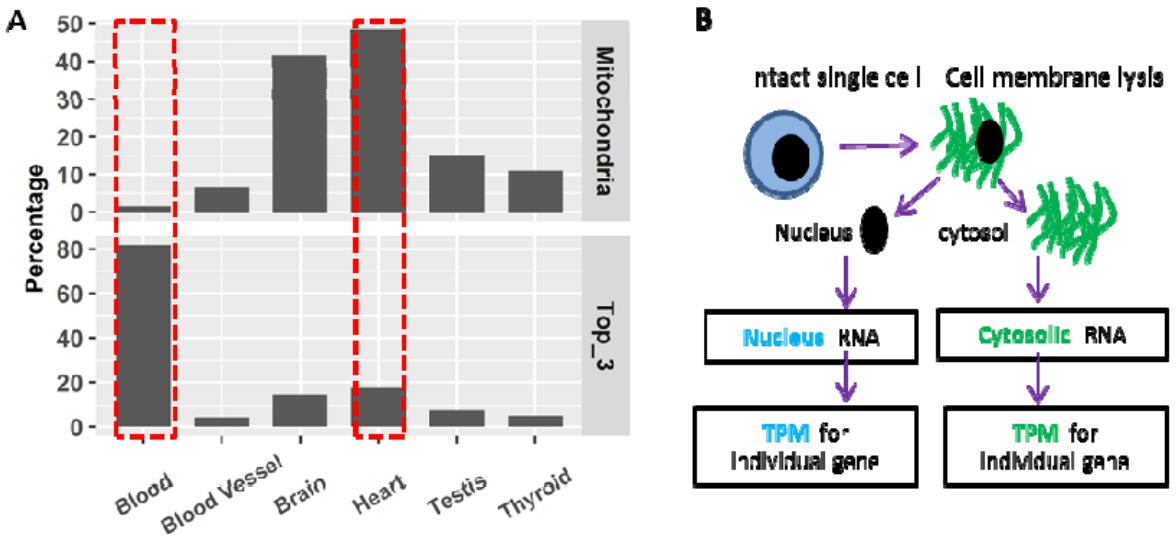

Figure 2. A) The percentages of transcripts from mitochondria, and the top three most abundant transcripts, in different tissue samples of the same subject (GTEX-N7MS) from the GTEx project. B) In cellular fractionation RNA sequencing, the nucleic and cytosolic RNA populations are very different, and thus TPM values are not directly comparable.

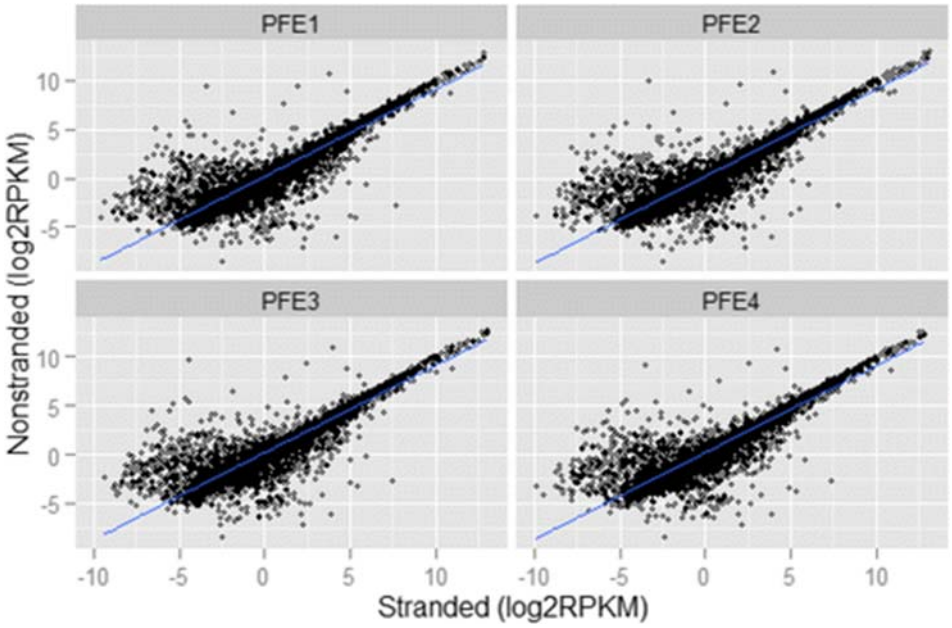

Figure 3. Scatter plots of gene expression profiles between stranded and non-stranded RNA-seq. For blood biological replicates PFE1, PFE2, PFE3, and PFE4, the scattering patterns are consistent. While the majority of genes are arrayed along the diagonal lines, there are still many genes whose expression levels are dramatically impacted by sequencing protocols. The x- and y-axis represent Log2(RPKM).

Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols

Shanrong Zhao, Zhan Ye and Robert Stanton

RNA published online April 13, 2020